CASE STUDY- 1

MATH 189/ MATH 289C

UCSD Winter 2017

Name	PID	Degree & Major
Himanshu Makhija	A09845605	B.S.- Math/CS
Hanna Goldman	A11436767	B.S.- Math/CS
Amey Paranjape	A53218045	M.S.- ECE
Shagun Gupta	A91068956	B.S.-BENG:Bioinformatics

Table of Contents

Objective

Our objective here is to answer the following questions by doing the statistical study of the given datasets:
What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

Data Collection

Source of Data

Data collected for our study is enlarged portion of the mentioned CHDS data. The data consists of all pregnancies that occurred between 1960 and 1967 among women in Kaiser Health Plan in Oakland, California.

Population of Interest: All women who delivered babies recently

Sample of Interest: The women in the study are all those that were enrolled in Kaiser Heath Plan, had obtained prenatal care in San Francisco area and delivered in any of the Kaiser hospitals in Northern California. Our study is comprised of 1236 babies: All boys, single births (no twins), all lived at least 28 days

Population to which results can be generalized: Adult women who will give birth to boys as only child and not twins.

Healthy Baby Parameters:

Healthy Gestation Range: 259 up to 294 days, Healthy Baby Body Weight: 88 up to 141 ounces. However the gestation range is not as important a factor as body weight. This is because despite having an abnormal gestation period, if the body weight is normal the baby can be considered to be healthy.

Both text-files: babies.txt and babies23.txt are essentially the same dataset. The dataset babies.txt is a more concise and compact version of the dataset, focusing primarily on baby bodyweight, gestation time and mother's height and weight. The more detailed dataset focuses on more parameters mentioned later. So for the smaller dataset we need to verify whether the body weight of the baby, which is an indicator of the babies health, is dependent in any way on whether the mother smokes or not. The key variables into question are "bwt" and "smoke". The other variables "gestation", "parity", "height", "weight" may come into the picture as confounding variables. For example, if the body weight of the baby is low and the gestation time is low as well, this can be due to the premature birth and not because of smoking in particular.

Hypothesis

Based on the observation of the dataset babies.txt, we hypothesize that birthweight decreases if the mother smoked during the pregnancy. Factors such as weight of the mother and parity do not have an effect on the birthweight of the baby by themselves. We also predict that the gestation period for the pregnancy would not effect the birthweight of the baby.

Classification of Variables

Variable	Description	Type of Variable
Bwt	Birth weight in ounces (999 unknown)	Numerical, Continuous
Gestation	Length of pregnancy in days (999 unknown)	Numerical, Discrete
Parity	0= first born, 9=unknown	Categorical, Normal
Age	mother's age in years	Numerical, Discrete
Height	mother's height in inches (99 unknown)	Numerical, Continuous
Weight	Mother's pre-pregnancy weight in pounds (999 unknown)	Numerical, Continuous
Smoke	Smoking status of mother (0=not now, 1=yes now, 9=unknown)	Categorical, Ordinal

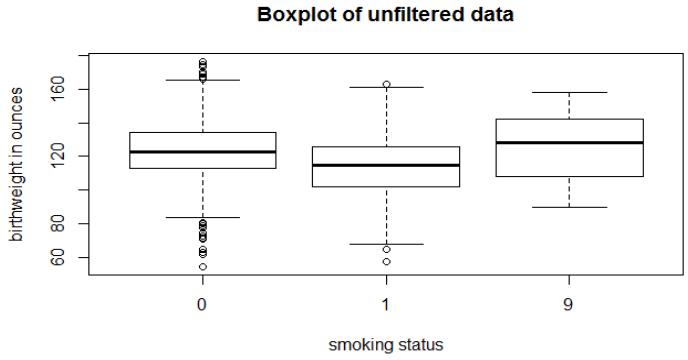
Table 1. Concise Dataset

Analysis

Summary of Data

The data set sampled on originally contained 1236 observations of mothers. Several mother's decided not to answer one or multiple of the questions in the study. These mothers were omitted from the analysis. This resulted in a total of 1174 observations. After this the data was observed and an extreme outlier in the set was noted, a mother having a baby with a gestational age of only 148 days. This data point would skew the potential findings and was omitted. This left the set with a total of 1173 observations. Figures 1 and 2 show box plots comparing the data before and after cleaning.

Since the goal of the study is to compare smoking mothers and nonsmoking mothers, the data was then separated into those two categories. The set of smoking mothers contained 459 observations, and the set of nonsmoking mothers had 715 mothers. Tables 2.3 and 2.4 show the summary statistics for the two sets.



B

Figure 1. Box Plot of uncleaned data for the set of mothers

As can be seen from this Figure 1, there are a greater number of data-points as outliers for non-smoking mothers versus that of the smoking mothers. The average birthweight from smoking mothers also comes out to be less than that from nonsmoking mothers as can be observed by the difference in height of the boxes and thus the difference of the mean lines. The data points for an unspecified smoking status of the mother do not show any outliers.

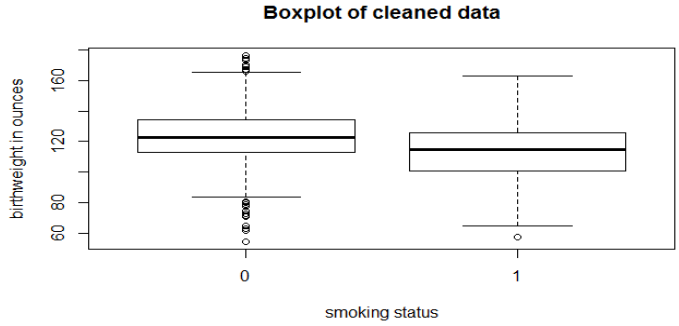


Figure 2. Box Plot of cleaned data for the set of mothers

Figure 2 clearly indicates the cleaning of outliers such as the smoke value of 9, which implied that the smoking status of the mother was unknown or not specified. All gestation periods of 999 days, age of 99 years, height of 999 and 99, weight of 999 ounces were removed. This can be indicated by the significant cleaning of outliers for the smoking status of 1 (mothers that smoked during their pregnancy).

	Birth Weight	Gestation	Parity	Age	Height	Weight	Smoke
Min	58.0	223.0	0.000	15.00	53.0	87.0	1
Median	115.0	278.0	0.000	26.00	64.0	125.0	1
Mean	113.8	277.9	0.257	26.74	64.1	126.9	1
Max	163.0	286.0	1.000	43.00	72.0	215.0	1
Standard Deviation	18.295	15.201	0.437	5.713	2.603	19.991	0

Table 2. Summary of Set of Smoking Mothers

	Birth Weight	Gestation	Parity	Age	Height	Weight	Smoke
Min	55.0	181.0	0.000	17.0	56.00	89.0	0
Median	123.0	281.0	0.000	27.0	64.00	126.0	0
Mean	123.1	280.1	0.266	27.5	64.01	129.5	0
Max	176.0	353.0	1.000	45.0	71.00	250.0	0
Standard Deviation	17.434	15.726	0.442	5.870	2.476	21.164	0

Table 3. Summary of Set of Nonsmoking Mothers

Relationship Between Gestational Age and Birth Weight

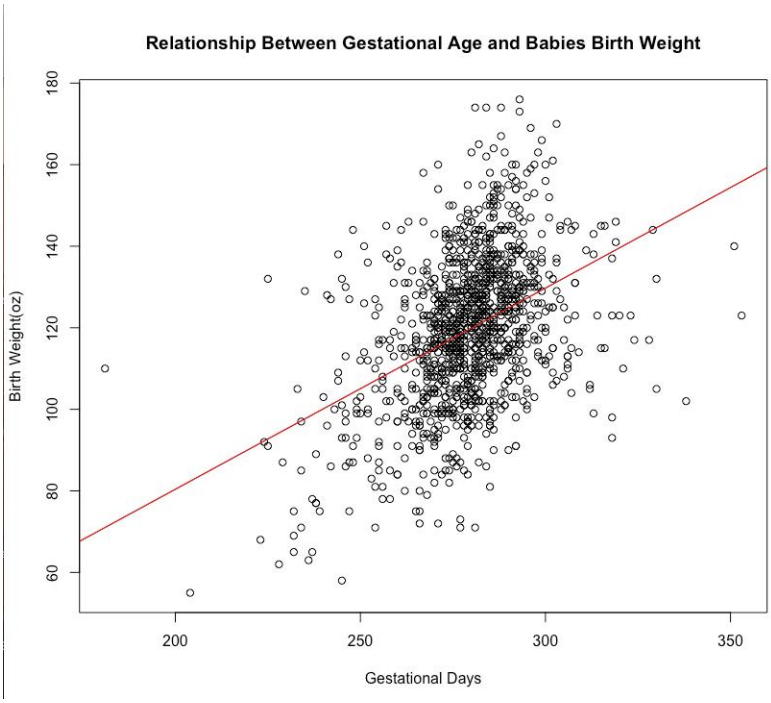


Figure 4. Scatterplot for Gestational Age vs Birth Weight

The scatterplot shows the locations of data-points in the XY-plane where X signifies Gestation time and Y signifies bodyweight. Typically, the healthy baby should fall in the normal range of both gestation time and body weight. The vertical line shows the minimum number of gestation period for a healthy baby, and the horizontal line shows the minimum body weight for a healthy baby. This is obvious as mothers who tend to have expected gestation time have healthier babies and when the gestation time is low, naturally the baby weight will be low.

Hence based on the scatterplot we can say: ***The "gestation" and the "bwt" are positively associated.***

Effect of Mothers BMI on Birth Weight

To determine whether smoking played a role in the weight of the baby, other factors needed to be considered to see if they had a significant effect as well. One potential factor was a mother's BMI (body mass index). The BMI was calculated for each mother from their given height and weight.(http://extoxnet.orst.edu/faqs/dietcancer/web2/twohowto.html)

This was then plotted against the given birth weight of the baby in Figure 5. A healthy BMI was decided to be between 19-25 and the average healthy birth weight of a baby was decided to be between 5.5-8.8lbs. Lines to determine these ranges were then added to the graph.

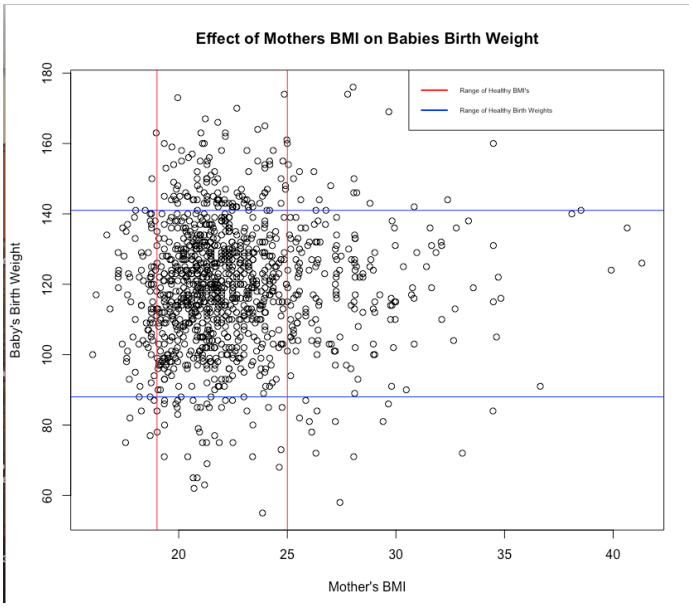


Figure 5. Scatterplot for Mother's BMI vs Birth Weight

To determine whether mother's BMI played a role in the weight of the baby, data points outside of the red lines were observed. If a mother's BMI had a role in the weight of the baby, then points outside of the range of healthy BMI would result in them to go further out of the range for healthy baby weights. This graph shows though that the points outside of the range of healthy BMI's tend to still stay in between the range of healthy birth weight's for babies and that the most variance is actually present within the range of healthy BMI's. Therefore it was determined that a mother's BMI did not play a significant role in impacting a baby's birth weight. This means that the variance in birth weights may be caused by another factor, potentially smoking.

Effects of Parity on Birthweight

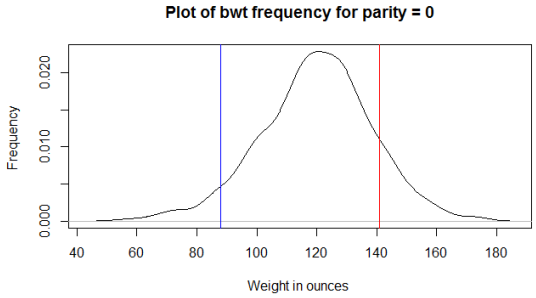


Figure 6. Frequency plot for parity=0

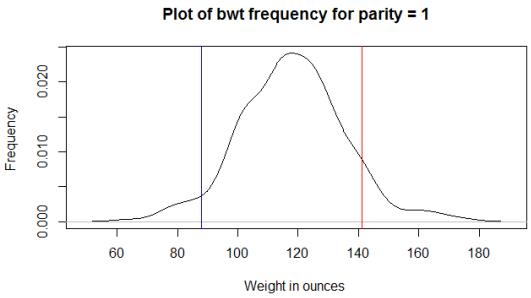


Figure 7. Frequency plot for parity=1

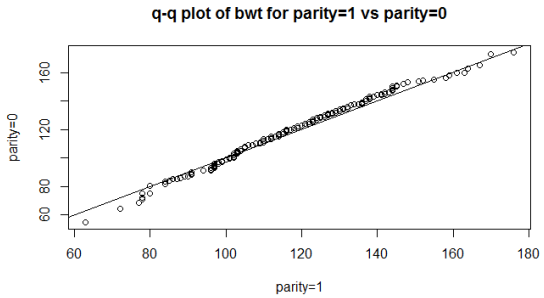


Figure 8. Q-Q plot for Parity

In the above frequency plots, it can be clearly seen that the distribution of birthweight is a normal distribution. The two colored lines indicate the thresholds of the healthy baby's birthweight. Also the trend normal distribution is confirmed by the Q-Q plot for parity. Though it is very clear distribution, there is nothing much to extract from this comparison. So there is no clear relationship between the parity and baby's birthweight.

Q-Q Plot: Observing the Distribution of Birthweights

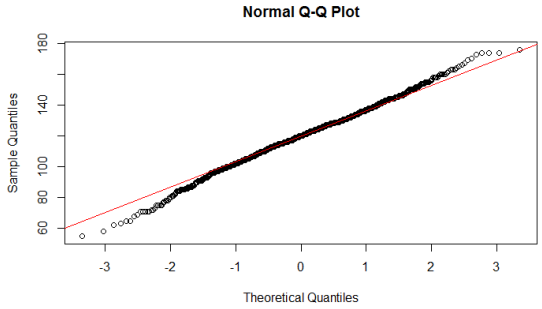


Figure 9. Birthweight for Cleaned Dataset

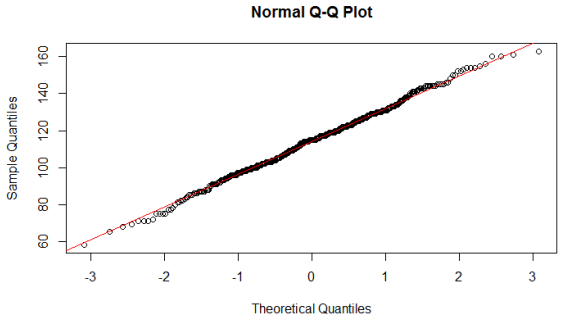


Figure 10. Birthweight (Smoking Mothers)

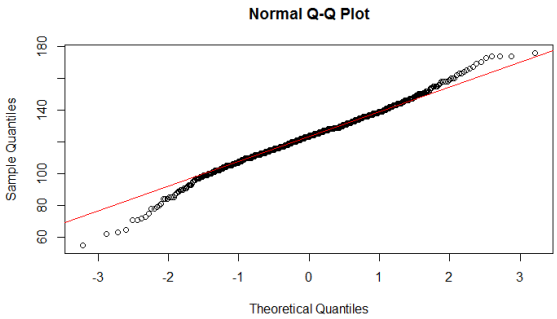


Figure 11. Birthweight (Non-Smoking Mothers)

The quantile-quantile plots indicate how close the data is to normal distribution. By the reference line we can clearly see that all the plots are close to normal distribution. The smoking mother's plots especially clearly follows the normal distribution curve even at the outliers.

Difference in Birth Weights for Babies of Nonsmokers vs Smokers

Next the birth weights for babies of nonsmokers and smokers was plotted to determine if there was a significant difference between the two distributions. A line was plotted to show the cut-off for what was deemed a "small" baby(<5.5lbs) and a line was plotted to show the median for the data set. Along with those lines, a curve to show a normal distribution was plotted to compare it with the distribution of the data.

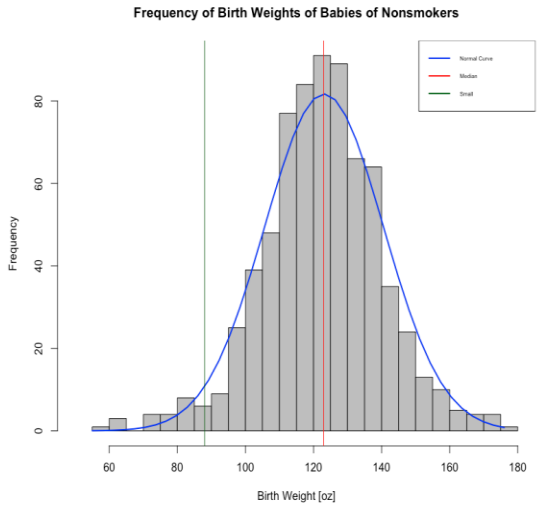


Figure 12. Birth Weights for Babies of Nonsmokers

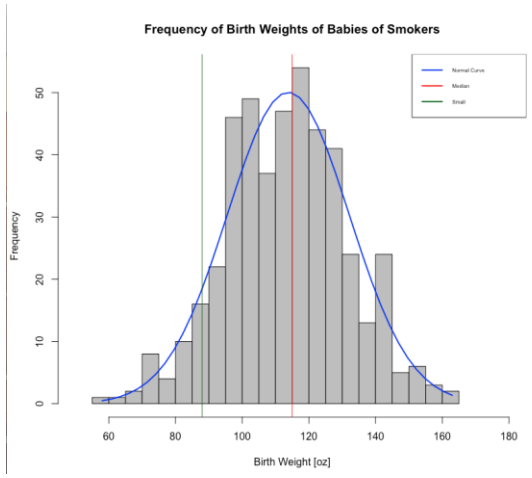


Figure 13. Birth Weights for Babies of Smokers

The median birth weight for babies of nonsmokers is 123 oz. compared to a median of 115 oz. for babies of smokers. The percent of babies considered “small” (less than 5.5 pounds) for smokers is 7.8% while the percent of babies considered “small” for nonsmokers is 2.9%.

The percent of babies of smokers that are considered "small" is larger than the percent for nonsmokers. This means that smoking may contribute to lower birth weights of babies.

Effects of Smoking on Gestation Period

Next, the number of gestational days for babies of nonsmokers and smokers was plotted to determine if there was a significant difference between the two distributions.

Lines were plotted on the graph to show the cutoff for what is considered a preterm(born before 37 weeks) baby and to show the range for the average number of gestational days(40-42 weeks). Along with those lines, a curve to show a normal distribution was plotted to compare it with the distribution of the data.

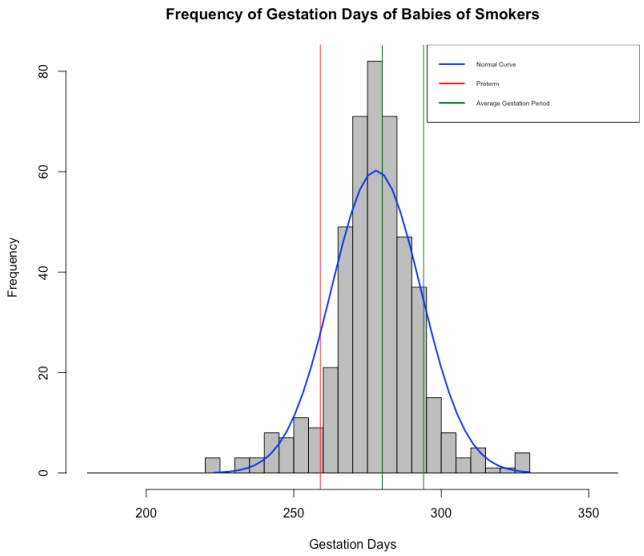


Figure 14. Gestation Days for Babies of Smokers

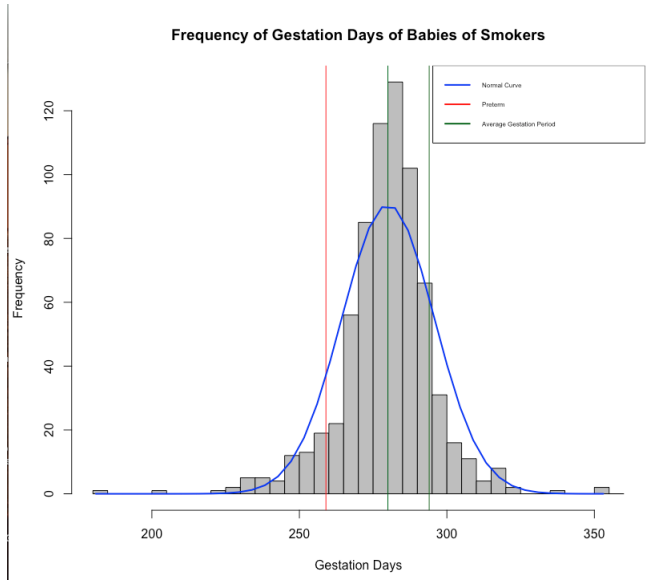


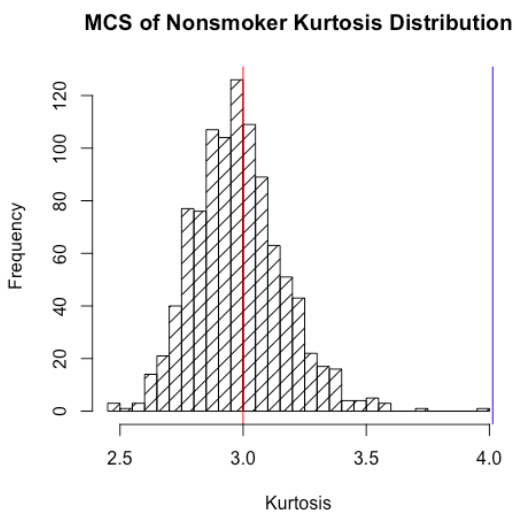
Figure 15. Gestation Days for Babies of Nonsmokers

It can be observed that the middle of the distribution for nonsmokers falls within the range of average gestational days while the middle of the distribution for smokers falls just under the range.

The percent of babies within the average gestational period for smokers is 30.7% while the percent of babies within the average for nonsmokers is 39.4%. The percent of babies considered preterm for smokers is 8.7% while the percent of babies considered preterm for nonsmokers is 7.7%.

Therefore it can be observed that on average, babies of smokers are born at a younger gestational age than babies of nonsmokers.

Confirming Distributions – MC on Kurtosis



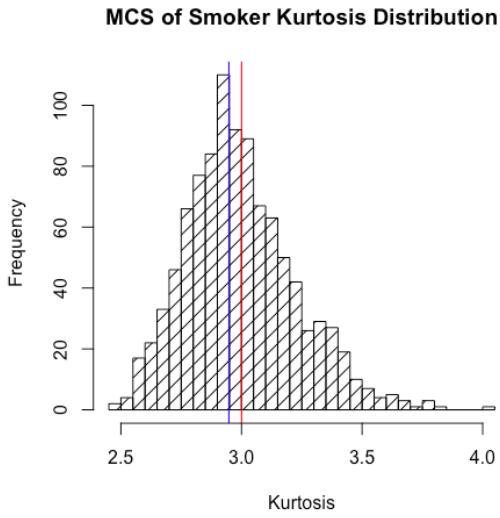


Figure 16. Kurtosis Distribution of Nonsmokers **Figure 17. Kurtosis Distribution of Smokers**

On the left histogram we are seeing the result of a simulation done on the kurtosis of 1000 normally distributed samples of size 715; each generated sample follows mean of 0 and variance 1. In red is the expected kurtosis value of a standard normal distribution, whereas the blue line is the kurtosis of the sample of only nonsmokers.

In contrast, the histogram on the right displays the distribution of the kurtosis of 1000 standard normal distribution samples of size 459, in accordance with the sample size of the smoker group. Notice the blue line, which represents the sample kurtosis, is relatively closer to the theoretical value.

The nonsmoker histogram displays the sample kurtosis as heavily right, which shows that it tends to deviate from the expected distribution of birthweight points in the tails of the sample distribution. The smoker group seems to follow the expected kurtosis of a standard normal distribution, with a slight skew left which would imply lighter-tails in the distribution of our birthweight for nonsmokers.

Conclusion

Using the complete datasets given we are trying to answer the question: What is the difference in weight between babies born to mothers who smoked during pregnancy and those who did not? Is this difference important to the health of the baby?

The concise dataset is a more useful for quick analysis to see the kind of association between the weights of the babies and maternal smoking. We clearly observe a positive association between the two variables. Although the scatter-plot is not as conclusive, the histogram and the frequency of incidence clearly show that smoking affects the body weight and the number of gestational days of the baby in a negative manner.

Other studies have shown that lower numbers of gestational days and lower birth weights of babies are strongly associated with mortality risk during the first year of birth and with developmental problems. (https://academic.oup.com/ije/article/30/6/1233/651751/On-the-importance-and-the-unimportance-of) Therefore since smoking leads to the lowering of these two factors, it appears that smoking has a negative effect on the health of the baby.

Citations/Formulas

[1] **Sample Mean**: The arithmetic average of the set of values.

$$\overline{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

[2] **Sample Standard Deviation:** The dispersion of the dataset as a whole.

[3] **Quartiles:** each of any set of values of a variate that divide a frequency distribution into equal groups, each containing the same fraction of the total population.

[4] **IQR:** It is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

IQR = Q3 – Q1

[6] **Quantile-Quantile function:** It specifies, for a given probability in the probability distribution of a random variable, the value at which the probability of the random variable is less than or equal to the given probability.

[7] **Normal distribution:** The area under the normal curve is equal to 1.0. Normal distributions are denser in the center and less dense in the tails. They are defined by two parameters, the mean (μ) and the standard deviation (σ). 68% of the area of a normal distribution is within one standard deviation of the mean.

[8] **Kurtosis moment generating function:** Moments give an indication of the shape of the distribution of a random variable. Skewness and kurtosis are measured by the following functions of the third and fourth central moment respectively:

the coefficient of skewness is given by

$$\gamma_1 = \frac{\mathrm{E}(X-\mu)^3}{\sigma^3} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}};$$

the coefficient of kurtosis is given by

$$\gamma_2 = \frac{\mathrm{E}(X-\mu)^4}{\sigma^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3.$$

Paper on factor (cadmium and lead content of maternal hair) correlated to both parity and birthweight.

http://www.tandfonline.com/doi/abs/10.1080/00039896.1981.10667628